EDA: exploratory data analysis

|  |  |
| --- | --- |
| Type of data | public and private, and each has its own problem.  Public data:   * isn’t always relevant * collected by gov and public agencies   Private.   * Sensitive to organisations * <https://github.com/awesomedata/awesome-public-datasets> * <https://data.gov.in/> * <https://github.com/datameet> * <https://www.youtube.com/watch?v=jf-SIOFUuEo> (correlation) |
| Data cleaning | Fix rows 🡪 fix missing values 🡪 standardise values 🡪 fix invalid values 🡪 filter data |
| Checklist for Fixing Rows | * Delete summary rows: Total, Subtotal rows * Delete incorrect rows: Header rows, Footer rows * Delete extra rows: Column number, indicators, Blank rows, Page No. |
| Checklist for Fixing Columns | * Merge columns for creating unique identifiers if needed: E.g. Merge State, City into Full address * Split columns for more data: Split address to get State and City to analyse each separately * Add column names: Add column names if missing * Rename columns consistently: Abbreviations, encoded columns * Delete columns: Delete unnecessary columns * Align misaligned columns: Dataset may have shifted columns |
|  |  |

* Data Sourcing 🡪 Cleaning 🡪 Univariate analysis 🡪 Bivariate analysis 🡪 Derived matrix

|  |  |
| --- | --- |
| Let us summarise how to deal with missing values: | * Set values as missing values: Identify values that indicate missing data, and yet are not recognised by the software as such, e.g treat blank strings, "NA", "XX", "999", etc. as missing. * Adding is good, exaggerating is bad: You should try to get information from reliable external sources as much as possible, but if you can’t, then it is better to keep missing values as such rather than exaggerating the existing rows/columns. * Delete rows, columns: Rows could be deleted if the number of missing values are significant in number, as this would not impact the analysis. Columns could be removed if the missing values are quite significant in number. * Fill partial missing values using business judgement: Missing time zone, century, etc. These values are easily identifiable. |
| Working with missing data and why outlier treatment is required | * <https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html> * <https://www.kdnuggets.com/2017/02/removing-outliers-standard-deviation-python.html> |
| Invalid values | * Encode unicode properly * Convert incorrect data types * Correct values that go beyond * Correct values not in the list * Correct wrong structure * Validate internal rules |

|  |  |
| --- | --- |
| Univariate analysis | |
| **Types of Variables** | * Categorical variables  1. Unordered:   type of loan taken by a person: (home, personal, auto etc)  organisation of a person: Sales, Marketing, HR etc.  power low distribution on log scale   1. Ordered: high-medium-low, jan-feb-march :      * Quantitative/Numeric variables: salary, number of bank accounts.   Plot a histogram or bar graph to get an insights |
|  | 1. **Metadata description** describes the data in a structured way. You should make it a habit of creating a metadata description for whatever data set you are working on. Not only will it serve as a reference point for you, it will also help other people understand the data better and save time. 2. **Distribution plots** reveal interesting insights about the data. You can observe various visible patterns in the plots and try to understand how they came to be. 3. **Summary metrics** are used to obtain a quantitative summary of the data. Not all metrics can be used everywhere. Thus, it is important to understand the data and then choose what metric to use to summarise the data. |
| Segment univariate analysis | * 1. Basis of segmentation   2. Comparison of averages   3. Comparison of other metrics |